Sistemas Avanzados de Producción con Python

Aplicaciones de Estadística, Machine Learning y Simulación en Ingeniería Industrial

Fredy Orjuela - Parcial I (TAXIS NYC) - 2025 II

Nombre Estudiante: _____

Código: _____

Instrucciones Generales

- Responda las preguntas de selección múltiple rellenando la casilla correspondiente en la **Tabla de Respuestas** al final del examen.
- Utilice **esfero de tinta negra o azul**. No se calificarán respuestas a lápiz.
- **No se admiten tachones, enmendaduras o uso de corrector** en la tabla de respuestas.
- Lea atentamente cada pregunta antes de contestar.

Enunciado: Este parcial se basa en el análisis de un conjunto de datos masivo de viajes en taxis amarillos de Nueva York. El dataset contiene información detallada sobre la duración, distancia y costos de millones de viajes. Las principales medidas de tendencia central y dispersión para las variables numéricas se presentan en la Tabla 1. Responda las preguntas basándose en la tabla, los gráficos proporcionados y los conceptos teóricos vistos en clase.

	NumPasajeros	DistanciaViaje	MontoTarifa	MontoPropina	MontoTotal	DuracionViaje
count	2,711,675.00	2,711,675.00	2,711,675.00	2,711,675.00	2,711,675.00	2,711,675.00
mean	1.35	3.29	18.30	3.47	27.26	14.92
std	0.84	12.32	16.46	3.78	21.06	11.67
min	1.00	0.01	2.80	0.00	4.00	1.02
25 %	1.00	1.01	8.60	1.08	15.48	7.25
50 %	1.00	1.70	12.80	2.80	20.15	11.65
75 %	1.00	3.13	19.80	4.20	28.56	18.67
max	6.00	15,400.32	199.99	422.70	453.55	119.97

Cuadro 1: Medidas de Tendencia Central y Dispersión para el Dataset de Taxis de NYC.

1. ¿Cuál es el propósito principal de usar 'git' y GitHub en un proyecto de ciencia de datos?

a) Ejecutar código más rápido.

b) Versionar el código, es decir, guardar un historial de cambios y facilitar la colaboración.

c) Solo sirve para almacenar imágenes.

d) Es una base de datos para guardar tablas.

2. La librería Pandas es fundamental en el análisis de datos porque su estructura principal para manejar tablas es el:

a) Array

b) Plot

c) Lista

d) DataFrame

3. De acuerdo a la Tabla 1, ¿cuál es la duración promedio (mean) de un viaje en taxi?

a) 11.65 minutos.

b) 18.67 minutos.

c) 14.92 minutos.

d) 3.29 minutos.

4. En el contexto de un algoritmo, una notación Big-O de O(n) indica que su tiempo de ejecución:

a) Es constante.

b) Crece de forma lineal con el tamaño de la entrada.

c) Crece de forma cuadrática.

d) Es el más lento posible.

5. La Prueba de Hipótesis arrojó un p-value de 2.89e-21. ¿Qué se concluye sobre la duración del viaje según el método de pago?

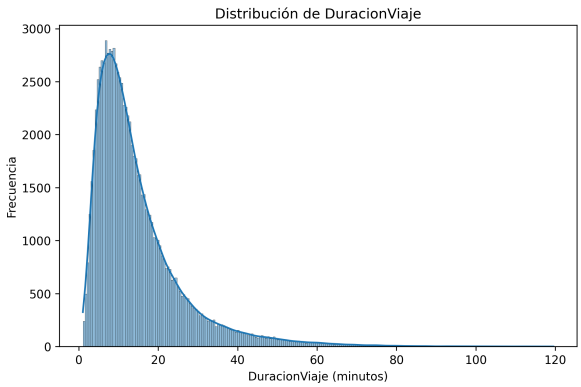
a) El p-value es alto, por lo que las medias son iguales.

b) El p-value es extremadamente bajo, por lo que se rechaza H_0, concluyendo que hay una diferencia significativa en la duración.

c) No se puede concluir nada con ese p-value.

d) El método de pago no afecta la duración.

6. Observe el histograma de la "Distribución de DuracionViaje".



¿Qué característica principal muestra la distribución?

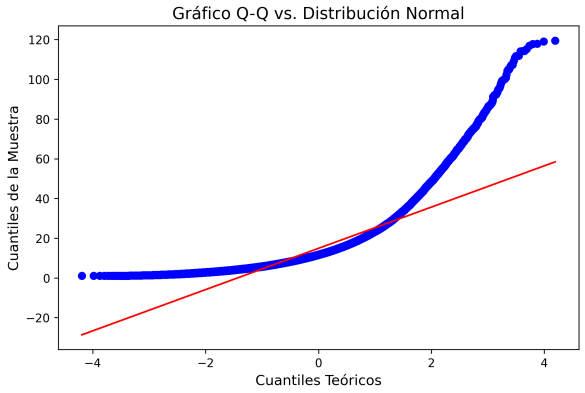
a) Los datos siguen una distribución normal.

b) La distribución está fuertemente sesgada a la derecha, indicando que la mayoría de los viajes son cortos.

c) La mayoría de los viajes duran más de 60 minutos.

d) Es una distribución uniforme.

7. El Gráfico Q-Q para la variable 'DuracionViaje' muestra que:



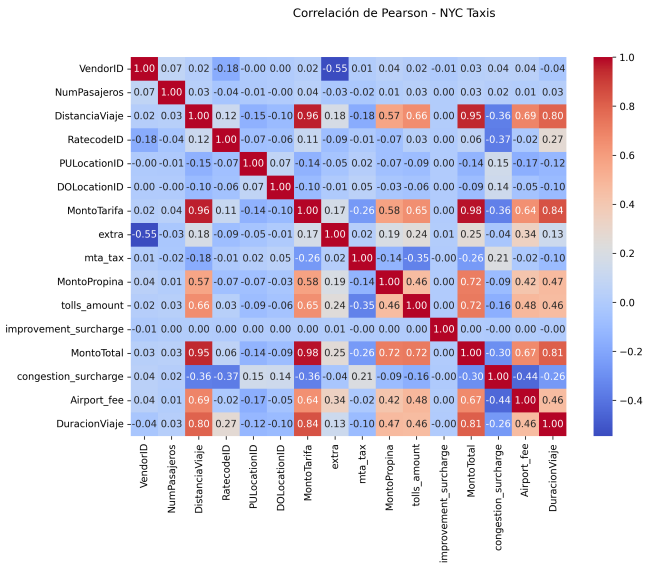
a) la variable sigue una distribución normal perfecta.

b) los datos se alejan significativamente de la normalidad, especialmente en la cola superior (viajes largos).

c) los datos se ajustan perfectamente a la línea roja.

d) el gráfico indica que no hay valores atípicos.

8. Analice la matriz de Correlación de Pearson.



¿Qué par de variables tiene la correlación lineal positiva más fuerte?

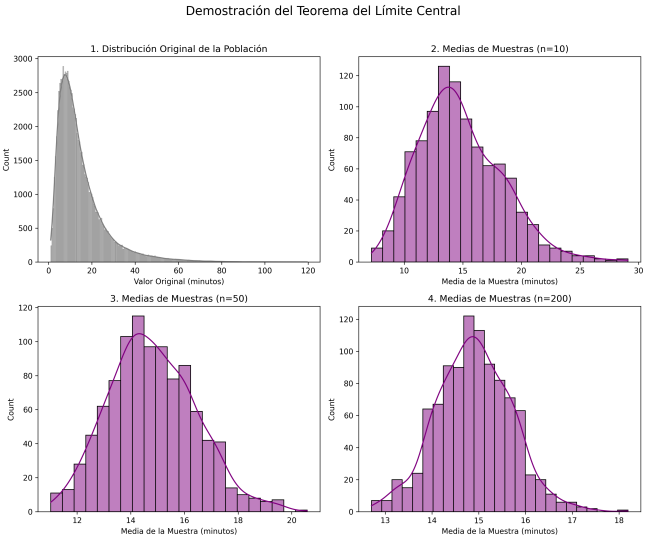
a) ‘DistanciaViaje‘ y ‘MontoTotal‘ (0.81).

b) ‘DuracionViaje‘ y ‘MontoTotal‘ (0.81).

c) ‘MontoTarifa‘ y ‘MontoTotal‘ (0.98).

d) ‘MontoPropina‘ y ‘MontoTotal‘ (0.74).

9. La imagen del Teorema del Límite Central demuestra que, aunque la ‘DuracionViaje‘ está muy sesgada...



a) la distribución de las medias muestrales también es sesgada.

b) la distribución de las medias muestrales tiende a la normalidad a medida que ’n’ aumenta.

c) la distribución original es la más precisa.

d) el tamaño de la muestra no afecta la forma.

10. En Python, ¿qué librería se utiliza principalmente para crear visualizaciones y gráficos?

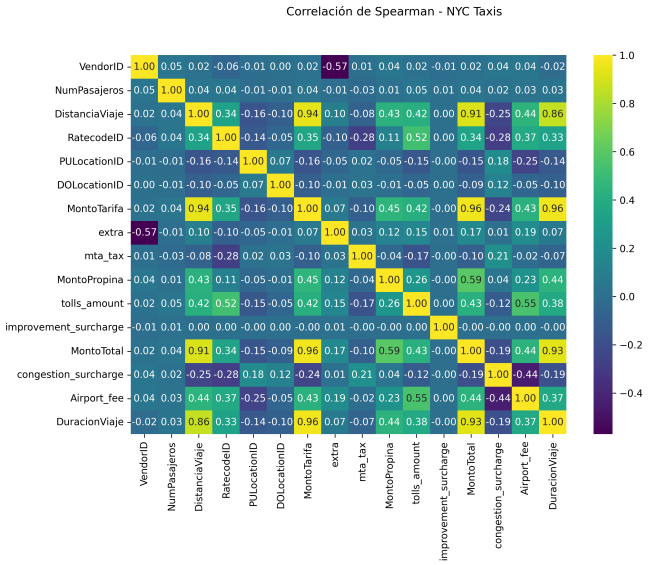
a) NumPy

b) Pandas

c) Matplotlib

d) Scikit-learn

11. Observe la matriz de Correlación de Spearman.



¿Cuál es la relación monotónica más fuerte que se observa?

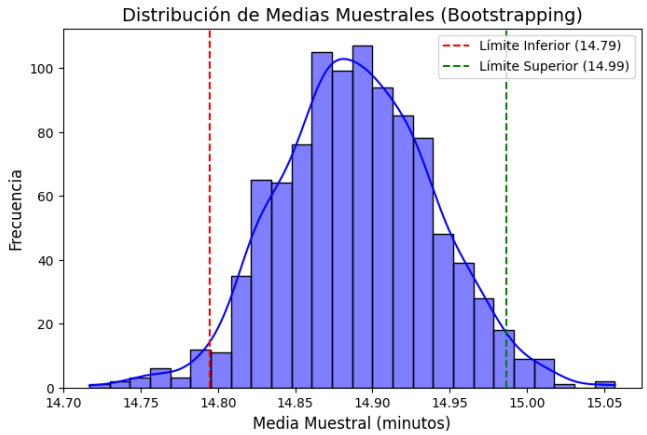
a) 'MontoTotal' y 'MontoTarifa' (0.96).

b) 'DuracionViaje' y 'DistanciaViaje' (0.86).

c) 'MontoTotal' y 'DistanciaViaje' (0.86).

d) 'MontoPropina' y 'MontoTotal' (0.60).

12. El gráfico de Bootstrapping arrojó un intervalo de confianza del 95 % de (14.85, 15.06) para la duración media del viaje.



¿Cómo se interpreta este resultado?

a) El 95

b) Se tiene un 95

c) La duración mínima fue 14.85 y la máxima 15.06.

d) La duración promedio de la muestra es 95 minutos.

13. ¿Qué es un commit en el contexto de Git y GitHub?

a) Un error en el código.

b) Una copia de seguridad completa del repositorio.

c) Un "punto de guardado"que registra los cambios realizados en los archivos, acompañado de un mensaje descriptivo.

d) El acto de descargar un repositorio.

14. ¿Para qué se utiliza principalmente la librería NumPy?

a) Para crear gráficos interactivos.

b) Para trabajar con tablas de datos (DataFrames).

c) Para realizar operaciones matemáticas eficientes sobre grandes listas de números (arrays).

d) Para crear modelos de Machine Learning.

15. La notación Big-O se utiliza para describir:

a) El número de líneas de un programa.

b) La cantidad de memoria RAM que usa un computador.

c) Cómo crece el tiempo de ejecución de un algoritmo a medida que aumenta el tamaño de la entrada.

d) La popularidad de un lenguaje de programación.

16. ¿Cuál es el propósito del archivo 'README.md' en un repositorio de GitHub?

a) Contener el código principal del programa.

b) Guardar las contraseñas y claves de acceso.

c) Mostrar una descripción general del proyecto, instrucciones de uso y otra información relevante.

d) Es un archivo temporal que se puede borrar.

17. La librería 'Seaborn' es una extensión de 'Matplotlib'. ¿Cuál es su principal ventaja?

a) Es más rápida para cálculos numéricos.

b) Permite crear gráficos estadísticos más atractivos y con menos código que Matplotlib.

c) Es la única librería para Machine Learning.

d) Sirve para conectar con bases de datos SQL.

18. ¿Qué significa clonar un repositorio de GitHub?

a) Crear una copia de seguridad en la nube.

b) Descargar una copia completa del proyecto y su historial a tu computador local.

c) Borrar el repositorio permanentemente.

d) Invitar a otros a colaborar.

19. Un algoritmo con complejidad O(log n) es considerado:

a) Muy ineficiente.

b) Lineal.

c) Cuadrático.

d) Muy eficiente, ya que el tiempo de ejecución crece muy lentamente.

20. ¿Qué librería contiene herramientas para realizar pruebas estadísticas como la prueba T o modelos de regresión lineal?

a) Pandas

b) Matplotlib

c) Statsmodels

d) NumPy

TABLA DE RESPUESTAS

Marque con una **X** la opción correcta.

Preg.	A	B	C	D
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				